Preprint
UCRL-ID- 155140

Stochastic Engine Convergence Diagnostics

R. Glaser

December 11, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

DISCLAIMER

Stochastic Engine Convergence Diagnostics

Ron Glaser 12/11/01

Introduction.

The stochastic engine uses a Markov Chain Monte Carlo (MCMC) sampling device to allow an analyst to construct a reasonable estimate of the state of nature that is consistent with observed data and modeling assumptions. The key engine output is a sample from the posterior distribution, which is the conditional probability distribution of the state of nature, given the data. In applications the state of nature may refer to a complicated, multi-attributed feature like the lithology map of a volume of earth, or to a particular related parameter of interest, say the centroid of the largest contiguous sub-region of specified lithology type. The posterior distribution, which we will call f, can be thought of as the best stochastic description of the state of nature that incorporates all pertinent physical and theoretical models as well as observed data. Characterization of the posterior distribution is the primary goal in the Bayesian statistical paradigm. In applications of the stochastic engine, however, analytical calculation of the posterior distribution is precluded, and only a sample drawn from the distribution is feasible. The engine's MCMC technique, which employs the Metropolis-Hastings [1, 2] algorithm, provides a sample in the form of a sequence (chain) of possible states of nature, $x^{(1)}$, $x^{(2)}$, …, $x^{(T)}$, …. The sequencing is motivated by consideration of comparative likelihoods of the data. Asymptotic results ensure that the sample ultimately spans the entire posterior distribution and reveals the actual state frequencies that characterize the posterior. In mathematical jargon, the sample is an ergodic Markov chain with stationary distribution f. What this means is that once the chain has gone a sufficient number of steps, T_0, the (unconditional) distribution of the state, $x^{(T)}$, at any step $T \geq T_0$ is the same (i.e., is "stationary"), and is the posterior distribution, f. We call T_0 the "burn-in" period. The MCMC process begins at a particular state, which is selected at random or

by design, according to the wish of the user of the engine. After the burn-in period, the chain has essentially forgotten where it started. Moreover, the sample $x^{(T_0)}, x^{(T_0+1)}, \ldots$ can be used for most purposes as a random sample from *f*, even though the $x^{(T_0+t)}$'s, because of Markovian dependency, are not independent. For example, averages involving $x^{(T_0)}, x^{(T_0+1)}, \ldots$ may have an approximate normal distribution.

The purpose of this note is to discuss the monitoring techniques currently in place in the stochastic engine software that addresses the issues of burn-in, stationarity, and normality. They are loosely termed "convergence diagnostics", in reference to the underlying Markov chains, which converge asymptotically to the desired posterior distribution.

The current engine has four convergence diagnostics, which will be considered separately in the ensuing sections.

1. A heuristic convergence diagnostic due to Gelman and Rubin [3] uses multiple parallel Markov chains to simultaneously estimate the burn-in period length T_0 and corroborate the claim of stationarity of the remaining samples.
2. A cumulative sum (cusum) plot due to Yu and Mykland [4] assesses the dependency between successive steps of a chain and as such is a measure of the speed of "mixing", i.e. how fast a chain steps through the posterior distribution.
3. Tests of stationarity of samples from the post burn-in portions of multiple chains due to Robert, Ryden, and Titterington [5] make use of Kolmogorov-Smirnov two sample statistics.
4. A test of normality of a selected mean based on post burn-in samples due to Robert, Ryden, and Titterington [5] makes use of a Kolmogorov-Smirnov one sample statistic.

No suite of convergence diagnostics is considered definitive at this time in the statistical literature. Promising alternative diagnostics that may be explored in future versions of the stochastic engine are discussed in Cowles and Carlin [6] and Robert [7].

The Gelman- Rubin Diagnostic.

Eventually a chain generated by the stochastic engine's MCMC process appears to have forgotten where it started and becomes usable as a sample from the posterior distribution. Ascertaining when this occurs is the goal of the Gelman-Rubin diagnostic.

A basic difficulty is introduced by the unknown character of the posterior distribution, which may contain several modes, or peaks, of relatively high value. An viable sample must visit each mode. Since the number of such modes is unknown, and because in sampling states of nature you only know where you have been, we can never be certain we have sampled sufficiently to have explored all the modes. This apparent impasse, however, is obviated by considering multiple independent (so-called parallel) chains with individual, well-dispersed starting points. Although the chains have different starting points, they share a common, but unknown, limiting distribution, the posterior *f*. The Gelman-Rubin diagnostic detects when the variability between the chains settles down to a value that is expected when the chains are all in a stationary condition of sampling a common distribution.

In many of our applications to date the parameter of interest is a multidimensional function of the state of nature, ψ. For example, for the Savannah River lithology problem, we split a cross section[1] of earth into two regions, upper and lower. We are interested in describing contiguous sub-regions

[1] The use of a two-dimensional cross section in this example rather than a three-dimensional volume is motivated by convenience not necessity. We have completed both algorithm development and coding for the three-dimensional case.

of specified lithology types, namely sand, clay, silt, and gravel. A sub-region is summarized by the triple $z = (z_1, z_2, z_3)$, where z_1 is the area, z_2 is the horizontal coordinate of the centroid, and z_3 is the vertical coordinate. By considering the largest contiguous sub-region for each of the two cross section halves and two of the lithology types, say clay and silt, the dimensionality of the parameter becomes p = 2x2x3 = 12. The multivariate Gelman-Rubin diagnostic, due to Brooks and Gelman [6], tracks the quantities R^p, det$V$, and det$W$, which are functions of the p-dimensional states of the parallel chains for a moving and expanding window of steps (called "iterations" by the authors). The window can be characterized by a single parameter n. For example, n = 50 refers to the window of length 50 iterations from iteration 51 through iteration 100, and in general, the window of size n considers each chain within the iteration sequence n+1, n+2, …, 2n. The p-dimensional matrix W estimates the within chain covariances for the window n, and the p-dimensional matrix B/n estimates the between chain covariances for the window. The pooled p-dimensional matrix

$$V = \frac{n-1}{n} W + \left[1 + \frac{1}{m}\right] B / n,$$

where m is the number of chains, is an estimate of the covariance matrix of the posterior distribution of the parameter of interest, ψ. As n increases, i.e. the window moves and expands, the influence of the starting points on the individual chains diminishes, and the following conditions begin to emerge:

- The within chain variation, summarized by the scalar quantity det$W$, stabilizes. Typically, det$W$ increases, as new areas of modality of the parameter space are explored by the chains, before settling to a limiting value once all areas are visited.
- The pooled chain variation, summarized by the scalar det$V$, stabilizes, a result of the combined effect of the difference between chains, characterized by B/n, becoming negligible and the within chain variation, characterized by W, stabilizing.
- The matrices V and W are "close".

Brooks and Gelman address the closeness issue by introducing a scalar measure of the distance between V and W:

$$R^p = \frac{n-1}{n} + \left[\frac{m+1}{m}\right]\lambda_1,$$

where λ_1 is the largest eigenvalue of the matrix $W^{-1}B/n$. As n increases, the distance between V and W diminishes, the eigenvalue λ_1 decreases to 0, and R^p approaches 1 from above. The Gelman-Rubin diagnostic, then, monitors R^p, det$V$, and det$W$, as a function of the window parameter n. For sufficiently large n, say $n \geq T_0$, the three conditions, R^p close to 1, det$W$ approximately constant, and det$V$ approximately constant, are satisfied. The nearness of R^p to 1 suggests burn-in has occurred by step T_0, in that the between chain variation is negligible (hence the starting points have been forgotten); stabilization of the determinants in turn provides evidence that samples within the window starting at iteration T_o+1are an adequate characterization of the stationary posterior distribution, since exploration of the parameter space has apparently succeeded in visiting all the modes.

Example plots of the statistic R^p and the determinants det$V$ and det$W$ as functions of n are shown in Figures 1 and 2 for the Savannah River lithology problem mentioned earlier in this section, but with dimension p = 8, based on analysis of the centroid but not the area. Four parallel chains were used in the simulation. The statistic R^8 seems to approach 1 and the determinants stabilize around $n = T_0 = 500$ iterations, the estimated burn-in length. The determinants are plotted in log scale because of the very large values that are associated with high dimensionality. Note that det$V$ always exceeds det$W$, and the two curves go up and down more or less together, ultimately converging. This is apparent from their definitions and the fact that the between chain factor diminishes with n. Actually, only one of the determinants need be plotted to investigate the stabilization condition, but both are presented as an additional check on the convergence of R^p.

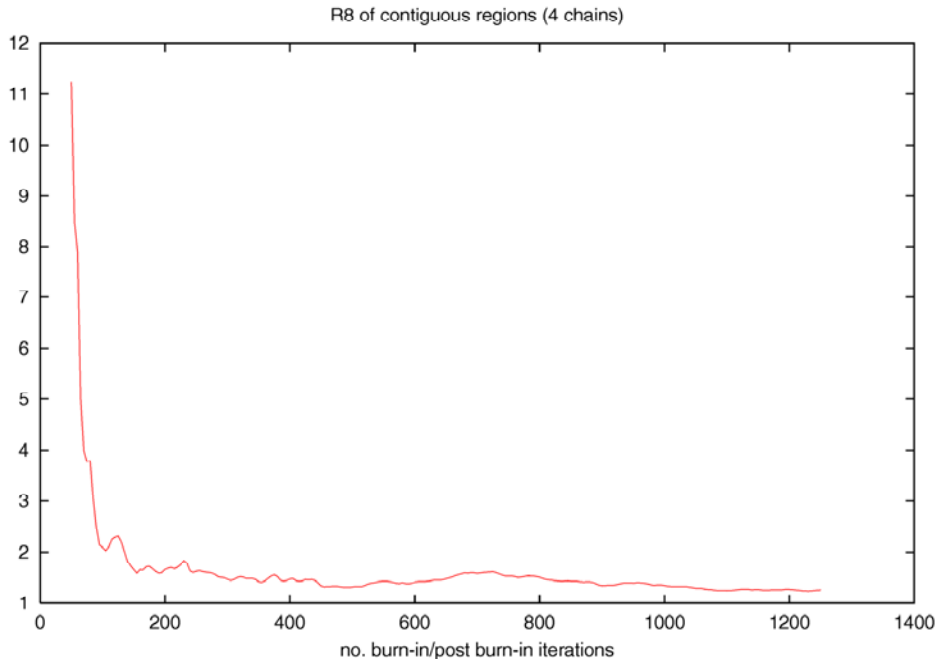

Figure 1. R^p for Savannah River Problem, p = 8, m = 4.

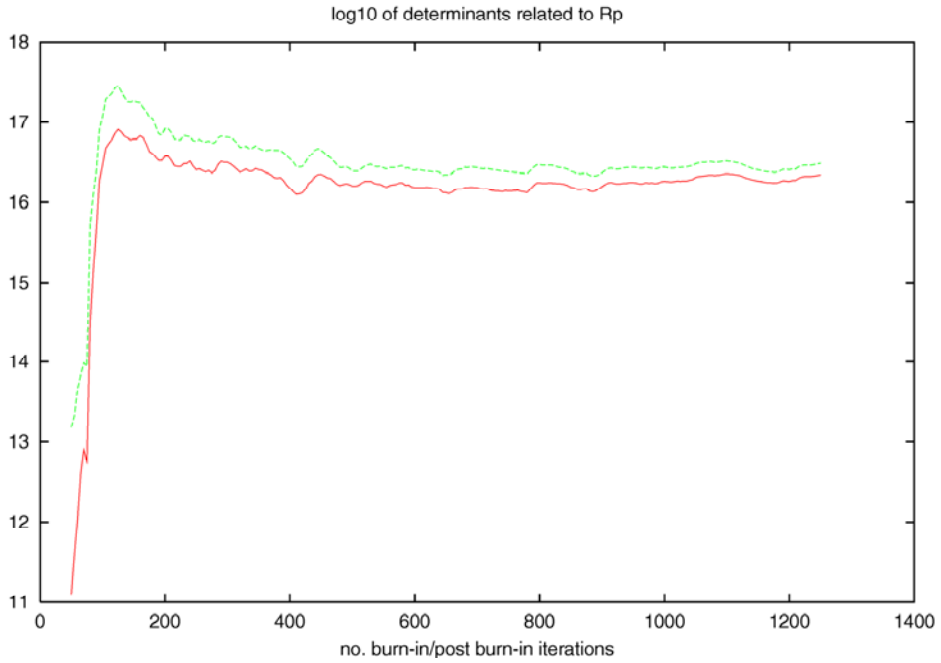

Figure 2. det$V$ and det$W$ for Savannah River Problem, p = 8, m = 4.

There is a measure of inefficiency and conservatism in the Gelman-Rubin approach due to the moving and expanding window parameterized by n. For each n, the first n iterations are discarded for each chain, and the next n iterations are used for the diagnostics R^p, det$V$, and det$W$. Thus the size of the sample discarded necessarily equals the size of the sample retained. The sample discarded must be large enough that each chain appears to have forgotten where it started, and the sample retained must be large enough that each chain appears to be sampling in its entirety the same distribution (the posterior). By using a single parameter, n, to handle both constraints simultaneously, we induce the conservatism of selecting a threshold T_0 that is the maximum of the sample sizes that would satisfy the individual constraints. Optimizing the process of selecting the burn-in length by treating the constraints separately is computationally unfeasible, so we accept the possible waste of samples caused by selecting too high a threshold. In practice, the simulation will generally be kept running after burn-in is diagnosed, say for a total of N iterations per chain counting from the start. Suppose there are m chains. If the Gelman-Rubin diagnostics suggest a burn-in period of length T_0, then a total of mT_0 samples are discarded, and $m(N - T_0)$ samples are available for analysis of properties of the state of nature.

The Cusum Plot.

The cumulated sum (cusum) plot monitors, for the MCMC sampling of a given chain, the partial sums

$$S_t = \sum_{j=T_0+1}^{t} \left[h(x^{(j)}) - \overline{h(x)} \right], \; t = T_0+1, \ldots, n,$$

where $h(x)$ is a scalar parameter of interest, say the y-coordinate (depth) of the centroid of the largest contiguous region of a particular lithology type within a

volume of earth, T_0 is the length of the burn-in period as estimated by the Gelman-Rubin diagnostic or a rough guess, and $\overline{h(x)}$ is the average value of $h(x)$ over the post burn-in steps T_0+1, …, n. The plot displays S_t versus t for the range $t = T_0, T_0+1, \ldots, n$, with $S_{T_0} \equiv 0$ and $S_n = 0$ by definition.

The cusum accumulates the differences between the value of a parameter at a given step and the overall average post burn-in value. It assesses the mixing behavior of the chain and correlation between the $x^{(t)}$'s. If the chain is slowly mixing the values of $h(x^{(t)})$ do not change much in a neighborhood of t, and the plot is smoother and wanders farther from zero than if the chain is faster mixing, in which case the plot may resemble Brownian motion.

The cusum is a subjective diagnostic that can be helpful in identifying samplers that are so slow mixing that alternative algorithms or parameterizations should be sought in order to more economically traverse the entire parameter space. Examples of the cusum are shown in Figures 3 and 4 for the dimension $p = 8$ Savannah River problem described in the previous section and for a contrived lithology problem of dimension $p = 2$ that deals with a sub-region (affectionately known as the "blob") of known size and lithology but unknown location. In each example the scalar parameter monitored is the depth of a contiguous region of specified lithology type, and the cusums of five parallel chains are plotted simultaneously. The Savannah River cusums of Figure 3 show faster mixing than those of Figure 4, the blob. Note that some chains are slower mixing than others for the plot window, evidence of a chain's hanging around a particular mode for an extended period.

Tests of Stationarity.

The Gelman-Rubin diagnostic is a heuristic which estimates a burn-in length T_0 while at the same time it asserts that samples after burn-in are from a stationary distribution, the posterior f. To test formally the contention of stationarity, which is

that $x^{(t)}$ and $x^{(t')}$ have the same distribution for arbitrary t and t' beyond burn-in, we adopt a batching approach, in which we divide the post burn-in samples into two

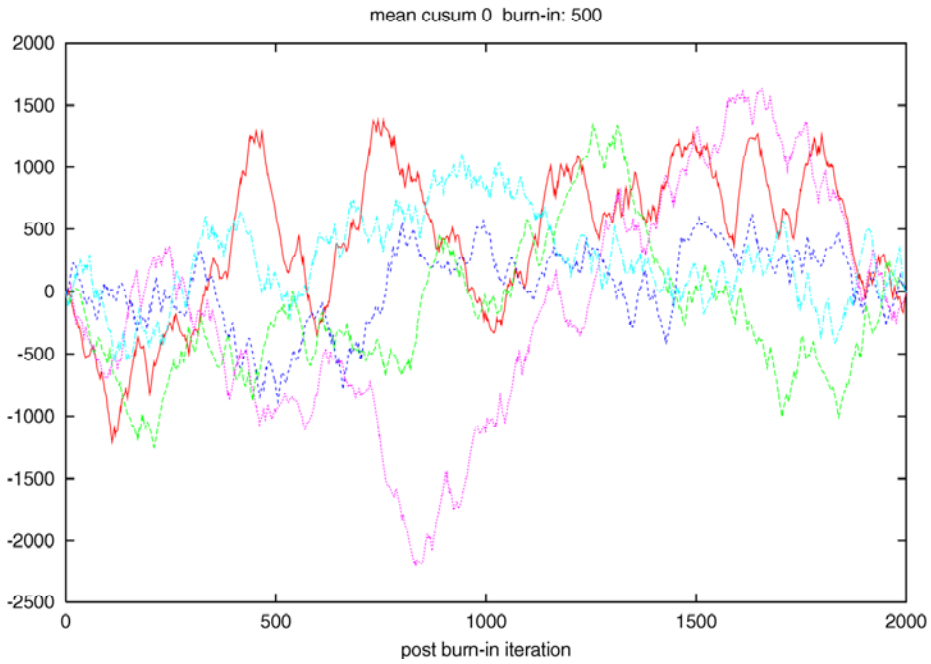

Figure 3. Cusums for Savannah River Problem, p = 8, m = 5.

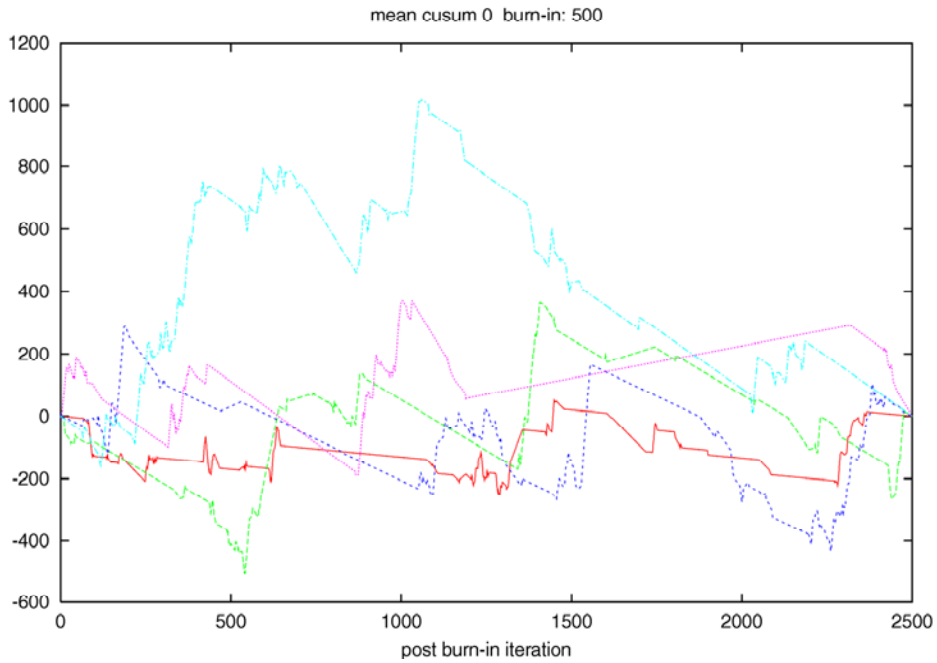

Figure 4. Cusums for Blob Problem, p = 2, m = 5.

halves and compare sub-samples from each half by means of a Kolmorogov-Smirnov test. Acceptance of the test is evidence of internal stationarity of the chain.

Specifically, consider a scalar parameter of interest $h(x)$, as in the cusum. Assume a burn-in length of T_0, obtained by the Gelman-Rubin diagnostic or a rough guess. Denote by N the total number of iterations counting from the start and $T = N - T_0$ the total number of post burn-in iterations. We compare the two sub-samples $h(x^{(T_0+1)}), \ldots, h(x^{(T_0+T/2)})$ and $h(x^{(T_0+T/2+1)}), \ldots, h(x^{(T_0+T)})$. In order to achieve approximately correct p-value levels with the classical two sample Kolmogorov-Smirnov (KS) test, we need to reduce the dependence created by the Markovian structure. Hence each sub-sample is itself sub-sampled, by retaining every r^{th} iteration, resulting in two samples each of size $T/2r$. The KS statistic is essentially the largest absolute difference between the two empirical cumulative distribution functions (cdfs). The stochastic engine's implementation of the KS test consists of a plot of the KS p-value as a function of $T/2$, the size of each pre-sub-sampled half. The p-value is the probability of obtaining a KS statistic value as large or larger than that recorded by the engine's simulation, if in fact each half is sampling from the same underlying distribution, i.e. there is stationarity. Therefore the p-value provides evidence that the chain internally exhibits stationarity. The higher the p-value the stronger the evidence of stationarity. A p-value below 0.05 is generally considered reason to question the assertion that the two halves are random samples from a common distribution. Figure 5 is a plot of KS p-values for the Savannah River $p = 8$ problem described in previous sections, with $h(x)$ being centroid depth. Five parallel chains are plotted simultaneously, and an autocorrelation adjustment of $r = 10$ is used for each chain. The plots show evidence of stationarity, but their highly oscillatory

nature indicates the effect on the empirical cdfs of locally attracting modes in the posterior distribution, which can cause a chain to hang up for a while.

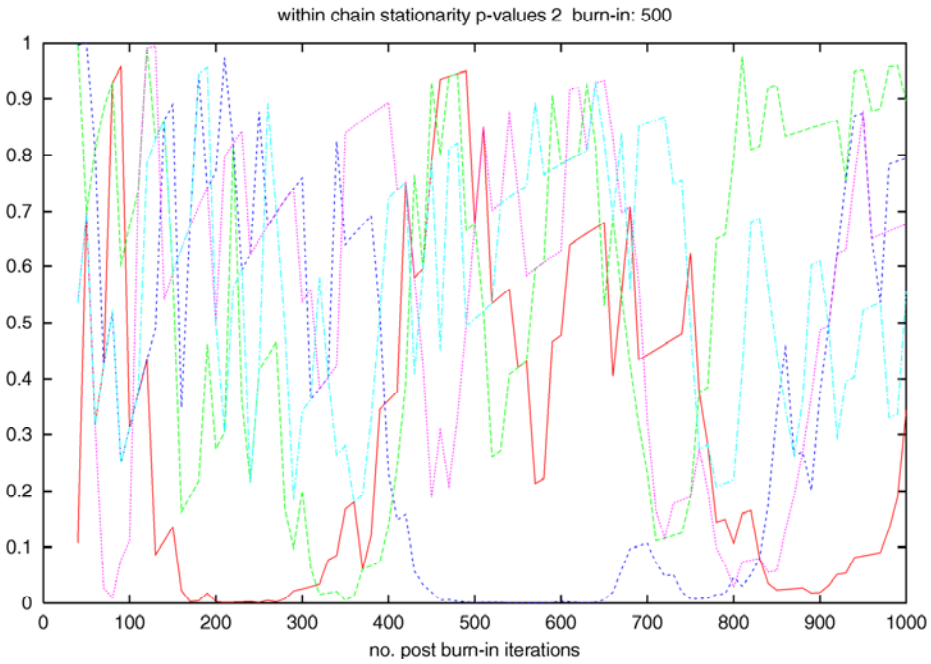

Figure 5. Within Chain Stationarity *p*-Values for Savannah River Problem, $p = 8$, $m = 5$.

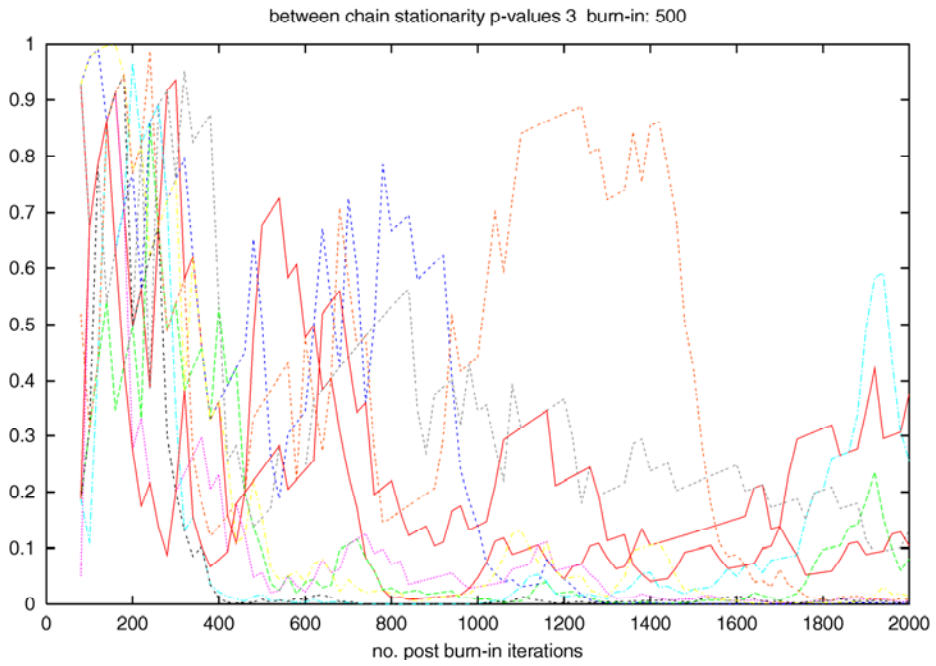

Figure 6. Between Chain Stationarity *p*-Values for Savannah River Problem, $p = 8$, $m = 5$.

After burn-in, the implication of stationarity is that $x^{(t,\,i)}$ and $x^{(t,\,j)}$ should have the same distribution for any $t \geq T_0$ and any two parallel chains i and j. This assertion can be tested formally by a Kolmogorov-Smirnov two sample test similar to the above for a single chain, in which we compare $h(x^{(T_0+1,i)}),\ldots,h(x^{(T_0+T,i)})$ and $h(x^{(T_0+1,j)}),\ldots,h(x^{(T_0+T,j)})$. Again reduce the Markovian dependence by retaining every r^{th} iteration so that each sub-sample has size T/r. Plots of p-values versus T for pairs of chains are shown in Figure 6 for the Savannah River problem of Figure 5. There is a curve for each of the 10 chain pairs. Evidence of stationarity is demonstrated for some of the pairs, with the kind of oscillatory character seen in Figure 5. However, there are some pairs, which have very small p-values throughout. This occurs when one chain is much slower mixing than the other for the given window of iterations. This common anomaly may be the reason Robert, Ryden, and Titterington stick with within chain stationarity tests and do not propose between chain stationarity tests.

Tests of Normality.

A goal of running the engine is to be able to summarize a selected feature of the state of nature, for instance the depth of a centroid of interest. The parameter may be estimated by a confidence interval obtained by application of the central limit theorem for MCMC algorithms. The central limit theorem asserts that the value of a parameter averaged over a large number of iterations after burn-in is approximately normally distributed. So we need to take some averages from our MCMC output and check for normality.

We use the notation of the previous sections: $h(x)$ is the MCMC measurement of a scalar parameter of interest, T_0 is the burn-in period, derived by the Gelman-Rubin diagnostic or a rough guess, m is the number of parallel chains, and T is the total number of iterations for a chain after burn-in. The issue of unknown correlation factors in the variance of $h(x)$ is treated by developing a sub-sampling device such that the sampling intervals grow with time, so that the dependence between sub-samples vanishes asymptotically. Specifically, define

$$\mu_{mT} = \frac{1}{mT} \sum_{j=1}^{m} \sum_{t=T_0+1}^{T_0+T} h(x^{(t,j)}),$$

and

$$v_{mT} = \frac{1}{mT} \sum_{j=1}^{m} \sum_{t=T_0+1}^{T_0+T} h^2(x^{(t,j)}) - \mu_{mT}^2,$$

which are consistent estimators of the mean and variance of $h(x)$, respectively, under the stationary distribution, f. For each chain j we introduce an independent sequence (u_{jk}) of independent integer-valued random variables defined by

$$u_{jk} \sim 1 + Poi(\nu_k),$$

where $Poi(\nu_k)$ denotes a Poisson random variable with mean ν_k, and the means satisfy $\nu_k = \nu k^d$ for some $\nu \geq 1$ and $d > 0$. The sums

$$t_{jk} = T_0 + u_{j1} + \cdots + u_{jk},\ k = 1, 2, \ldots$$

are then used as sampling instants for sub-sampling $h(x)$. The number of sampling instants having occurred up to T is

$$N_T = \sum_{j=1}^{m} k_j ,$$

where

$$k_j = sup\{k : t_{jk} \le T\}.$$

The central limit theorem then asserts that

$$Z_{mT} = \frac{1}{\sqrt{N_T v_{mT}}} \sum_{j=1}^{m} \sum_{k=1}^{k_j} \left[h(x^{(t_{jk}, j)}) - \mu_{mT} \right]$$

is approximately distributed as a standard normal random variable for sufficiently large N_T.

To test normality in the stochastic engine we construct a sequence of means of $h(x)$, with sampling instants developed as above, but with care taken that each mean has the same sample size n, where n is large enough that normality should be expected. In our applications, n = 40 has been used. The first mean in the sequence is that based on selection of $T = T_1$ such that $N_T = n$, i.e. T_1 is the minimum number of iterations for which a total of n sampling instants can be extracted from the m parallel chains. The sampling instants for chain j are t_{j1} through $t_{jk_{j1}}$, where

$$k_{j1} = \sup\{k : t_{jk} \le T_1\}$$

is the number of sampling instants devoted to chain j, and of course

$$\sum_{j=1}^{m} k_{j1} = n .$$

The second mean uses $T = T_2$ such that $N_T = 2n$, with the mean defined over the second set of n sampling instants: for chain j the sampling instants are $t_{jk_{j1}+1}$ through $t_{jk_{j2}}$, where

$$k_{j2} = sup\{k : t_{jk} \le T_2\},$$

and necessarily

$$\sum_{j=1}^{m} k_{j2} = 2n .$$

And so the process goes with the accumulation of n additional sampling instants at each stage. For the q^{th} stage we have $T = T_q$ such that $N_T = qn$, and the q^{th}

mean is defined over the q^{th} set of n sampling instants: for chain j the sampling instants are $t_{jk_{jq-1}+1}$ through $t_{jk_{jq}}$, where

$$k_{jq} = sup\{k : t_{jk} \le T_q\},$$

and

$$\sum_{j=1}^{m} k_{jq} = qn.$$

The random variables that are standardized versions of the sequence of means,

$$Z_{mT_q} = \frac{1}{\sqrt{nv_{mT_q}}} \sum_{j=1}^{m} \sum_{k=k_{jq-1}+1}^{k_{jq}} \left[h(x^{(t_{jk}, j)}) - \mu_{mT_q} \right], \; q = 1, 2, \ldots$$

may be accumulated and tested for normality by a one sample Kolmogorov-Smirnov (KS) statistic, which is essentially the maximum absolute difference between the empirical cdf of the sequence $\left(Z_{mT_q}\right)$ and the standard normal cdf.

The stochastic engine's implementation of the one sample KS test of normality consists of a plot of KS p-values as they become available in the sampling process. Specifically, once the Q standardized means $Z_{mT_1}, \ldots, Z_{mT_Q}$ are generated, the KS one sample statistic (with respect to sample size Q) based on the empirical cdf of these values and the standard normal cdf is calculated, and the corresponding p-value is calculated and plotted versus mT_Q, the total number of post burn-in samples associated with this collection of means. The p-value is the probability of obtaining a KS statistic value as large or larger than that recorded by the engine's simulation, if in fact there is standard normality of the standardized means. The higher the p-value the stronger the evidence of normality. A p-value below 0.05 is generally considered reason to question the validity of the assertion of normality. Figure 7 displays KS one sample p-values for the Savannah River p = 8 five chain problem depicted in previous figures. There is strong evidence of normality. In the simulation means of size n = 40 were generated, with sub-sampling based on a Poisson generator having ν = 10 and d = 0. The limiting choice of d was made for the sake of economy and also

because small values of *d* do not make a perceptible difference in the case of a few thousand iterations.

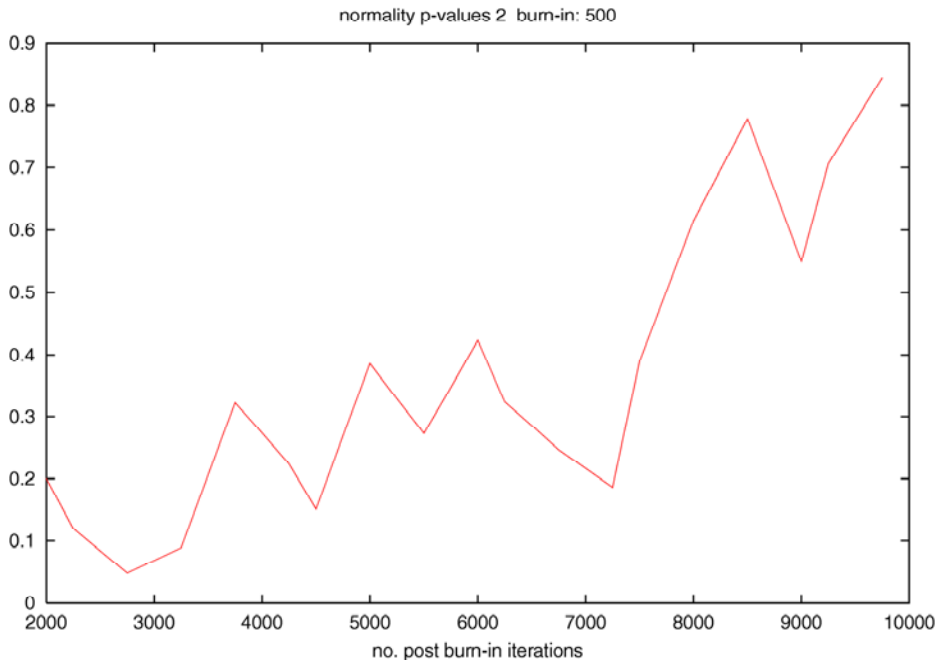

Figure 7. Normality *p*-Values for Savannah River Problem, $p = 8$, $m = 5$.

References.

1. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), “Equations of State Calculations by Fast Computing Machines”, Journal of Chemical Physics, 21, 1087-1091.

2. Hastings, W. K. (1970), “Monte Carlo Sampling Methods Using Markov Chains and Their Applications”, Biometrika, 57, 97-109.

3. Gelman, A., and Rubin, D. B. (1992), “Inference From Iterative Simulation Using Multiple Sequences”, Statistical Science, 7, 457-511.

4. Yu, B., and Mykland, P. (1994), “Looking at Markov Samplers Through Cusum Path Plots: A Simple Diagnostic Idea”, Technical Report 413, University of California at Berkeley, Dept. of Statistics.

5. Robert, C. P., Ryden, T., and Titterington, D. M. (1999), “Convergence Controls for MCMC Algorithms, With Applications to Hidden Markov Chains”, Journal of Statistical Computation and Simulation, 64, 327-355.

6. Cowles, M. K., and Carlin, B. C. (1996), “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review”, Journal of the American Statistical Association, 91, 883-904.

7. Robert, C. P. (1998), Discretization and MCMC Convergence Assessment, New York: Springer.

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551